Chapter 1
Dealing with Missing Depth: Recent Advances in Depth Image Completion and Estimation

Amir Atapour-Abarghouei and Toby P. Breckon

Abstract Even though obtaining 3D information has received significant attention in scene capture systems in recent years, there are currently numerous challenges within scene depth estimation which is one of the fundamental parts of any 3D vision system focusing on RGB-D images. This has lead to the creation of an area of research where the goal is to complete the missing 3D information post capture. In many downstream applications, incomplete scene depth is of limited value and thus techniques are required to *fill the holes* that exist in terms of both missing depth and colour scene information. An analogous problem exists within the scope of scene filling post object removal in the same context. Although considerable research has resulted in notable progress in the synthetic expansion or reconstruction of missing colour scene information in both statistical and structural forms, work on the plausible completion of missing scene depth is contrastingly limited. Furthermore, recent advances in machine learning using deep neural networks have enabled complete depth estimation in a monocular or stereo framework circumnavigating the need for any completion post-processing hence increasing both efficiency and functionality. In this chapter, a brief overview of the advances in the state-of-the-art approaches within RGB-D completion is presented whilst noting related solutions in the space of traditional texture synthesis and colour image completion for hole filling. Recent advances in employing learning-based techniques for this and related depth estimation tasks are also explored and presented.

Amir Atapour-Abarghouei
Department of Computer Science, Durham University, Durham, UK, e-mail: amir.atapour-abarghouei@durham.ac.uk

Toby P. Breckon
Departments of Engineering & Computer Science, Durham University, Durham, UK, e-mail: toby.breckon@durham.ac.uk

1.1 Introduction

Three dimensional scene understanding has received increasing attention within the research community in recent years due to its ever-growing applicability and wide-spread use in real-world scenarios such as security systems, manufacturing and future vehicle autonomy. A number of limitations pertaining to environmental conditions, inter-object occlusion and sensor capabilities still remain despite the extensive recent work and many promising accomplishments of 3D sensing technologies [33, 134, 149, 158]. It is due to these challenges that a novel area of research has emerged mostly focusing on refining and completing missing scene depth to increase the quality of the depth information for better downstream applicability.

Although traditional RGB image inpainting and texture synthesis approaches have been previously utilised to address scene depth completion [7, 39, 64], challenges regarding efficiency, depth continuity, surface relief, and local feature preservation have hindered flawless operation against high expectations of plausibility and accuracy in 3D images [4]. In this vein, this chapter provides a brief overview of the recent advances in scene depth completion, covering commonly-used approaches designed to refine depth images acquired through imperfect means.

Moreover, recent progress in the area of monocular depth estimation [6, 44, 55, 152] has lead to a cheap and innovative alternative to completely replace other more expensive and performance-limited depth sensing approaches such as stereo correspondence [129], structure from motion [27, 41] and depth from shading and light diffusion [1, 132] among others. Apart from computationally intensive demands and careful calibration requirements, these conventional depth sensing techniques suffer from a variety of quality issues including depth inhomogeneity, missing or invalid values and alike, which is why the need for depth completion and refinement in post processing arises in the first place.

As a result, generating complete scene depth from a single image using a learning-based approach can be of significant value. Consequently, a small portion of this chapter is dedicated to covering the state-of-the-art monocular depth estimation techniques capable of producing complete depth which would eliminate any need for depth completion or refinement.

1.2 Missing Depth

As explained in the previous chapter, different depth sensing approaches can lead to various issues within the acquired scene depth, which in turn make depth completion and refinement an important post processing step.

Passive scene sensing approaches such as stereo correspondence [129] have long been established as a reliable method of dense depth acquisition. Although stereo imaging is well-equipped to estimate depth where highly granular texture is present, even the smallest of issues in calibration and synchronization can lead to noisy, invalid or missing depth values. Additionally, missing values are prevalent in sections

Fig. 1.1: Examples of depth acquired via stereo correspondence (top), structured light device (bottom left) and time-of-flight camera (bottom right). **RGB:** colour image; **D:** depth image; **H:** hole mask indication missing depth values.

of the scene that contain occluded regions (i.e. groups of pixels that are seen in one image but not the other), featureless surfaces, sparse information for a scene object such as shrubbery, unclear object boundaries, very distant objects and alike. Such issues can be seen in Figure 1.1 (top), wherein the binary mask marks where the missing depth values are in a disparity image calculated via a stereo correspondence algorithm [65].

On the other hand, consumer devices such as structured light and time-of-flight cameras are active range sensors that are more-widely utilized for a variety of purposes due to their low cost and wide availability in the commercial market with factory calibration settings [14, 23, 46].

However, due to a number of shortcomings such as external illumination interference [23], ambient light saturation [46], inaccurate light pattern detection in the presence of motion [125] and active light path error caused by reflective surfaces or occlusion [126], consumer structured light devices can result in missing depth or noisy values that are best handled by removal and subsequent filling. An example of such a depth image and its missing values can be seen in Figure 1.1 (bottom left). Time-of-flight cameras can also suffer from complications detrimental to output deployment due to issues such as external illumination interference [123], light scattering caused by semi-transparent surfaces [59, 72] and depth offset for non-reflective objects [96]. Such issues are exemplified in Figure 1.1 (bottom right).

Completing depth images, captured through these active or passive depth sensing technologies, can lead to significant performance boost in any 3D vision application even though many current systems simply cope with challenges created by noisy and incomplete depth images without any post processing. In the next section, we will focus on various approaches to the problem of image completion in the context of RGB-D imagery.

1.3 RGB-D Completion

While object removal, inpainting and surface completion [2, 15, 17–20, 36, 43, 133] has been a long-standing problem addressed within the literature in the past few

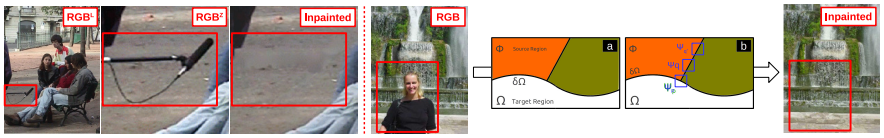

Fig. 1.2: **Left:** results of [15]. The foreground microphone has been removed and inpainted, but the texture is not accurate, leading to a perception of blurring. **Right:** an example of the results and process of exemplar-based inpainting [36].

decades, depth completion is a relatively new area of research with its own challenges and limitations. However, scene depth is still represented and processed in the form of images, and some researchers still directly apply classical RGB image inpainting methods to depth images or use depth completion approaches heavily inspired by RGB completion techniques. Consequently, an overview of image inpainting within the context of scene colour image (RGB) can be beneficial for a better understanding of the multi-facet subject of depth filling. In the following section, relevant image inpainting methods are briefly discussed before moving on to a more detailed description of the depth completion literature.

1.3.1 RGB Image Inpainting

Inpainting deals with the issue of a plausibly completing a target region within the image often created as a result of removing a certain portion of the scene. Early image inpainting approaches attempted to smoothly propagate the isophotes (lines within the image with similar intensity values) into this target area. However, most of these approaches [15, 133] tend to ignore an important aspect significant to an observer's sense of plausibility, which is the high-frequency spatial components of the image or texture. Consequently, later inpainting techniques began to incorporate ideas from the field of texture synthesis (in which the objective is to generate a large texture region given a smaller sample of texture without visible artefacts of repetition within the larger region [42, 43, 118]) into the inpainting process to compensate for the lack of texture commonly found in the target region post completion [2, 36, 79] (exemplar-based inpainting).

In one of the most seminal works on image inpainting [15], the problem is addressed using higher-order partial differential equations and anisotropic diffusion to propagate pixel values along isophote directions (Figure 1.2). The approach demonstrated remarkable progress in the area at the time but more importantly, it contained a set of guidelines for image inpainting created after extensive consultations with scene composition experts, which have now standardised the functionalities of an inpainting algorithm. These remain highly relevant even in depth completion:

- **1:** upon completion of the inpainting process, the target region must be consistent with the known region of the image to preserve global continuity.

- **2:** the structures present within the known region must be propagated and linked into the target region.
- **3:** the structures formed within the target region must be filled with colours consistent with the known region.
- **4:** texture must be added into the target region after or during the inpainting process.

Improved inpainting approaches were subsequently proposed employing a variety of solutions including the fast marching method [133], Total Variational (TV) models [28, 121], and exemplar-based techniques [16, 36]. In one such approach, [36] follows traditional exemplar-based texture synthesis methods [43] by prioritizing the order of filling based on the strength of the gradient along the target region boundary. Although [36] is not the first to carry out inpainting via exemplar-based synthesis [16], previous approaches are all lacking in either structure propagation or defining a suitable filling order that could prevent the introduction of blurring or distortion in shapes and structures. This exemplar-based method [36] is not only capable of handling two-dimensional texture but can plausibly propagate linear structures within the image. An example of the results of this method is seen in Figure 1.2 (right), in which water texture has been plausibly synthesized after the person is removed from the image. However, this approach cannot cope with curved structures and is heavily dependent on the existence of similar pixel neighbourhoods in the known region for plausible completion. Even though the approach relies on fine reflectance texture within the image to prioritize patches and can fail when dealing with large objects in more smooth depth images (Figure 1.3- left), it has been a great step towards focusing on granular texture within the image completion literature.

Other image completion techniques have also been proposed that would address different challenges in the inpainting process. For instance, certain methods use schemes such as reformulating the problem as metric labelling [85], energy minimization [12, 140], Markov Random Field models with labels assigned to patches [83], models represented as an optimal graph labelling problem, where the shift-map (the relative shift of every pixel in the output from its source in the input) represents the selected label and is solved by graph cuts [119], and the use of *Laplacian pyramids* [91] instead of the gradient operator in a patch correspondence search framework due to the advantageous qualities of Laplacian pyramids, such as isotropy, rotation invariance, and lighter computation. There have also been attempts to complete images in an exemplar-based framework using external databases of semantically similar images [60, 141] (Figure 1.3 - right).

Deep neural networks have recently revolutionized the state of the art in many computer vision tasks such as image stylization [52, 54, 76, 80], super-resolution [111, 138] and colourization [156]. Image completion has also seen its fair share of progress using such techniques. In [113], an approach is proposed that is capable of predicting missing regions in an RGB image via adversarial training of a generative model [56]. In a related work, [150] utilizes an analogous framework with similar loss functions to map the input image with missing or corrupted regions to a latent vector, which in turn is passed through their generator network that recovers the target content. The approach in [146] proposes a joint optimization framework

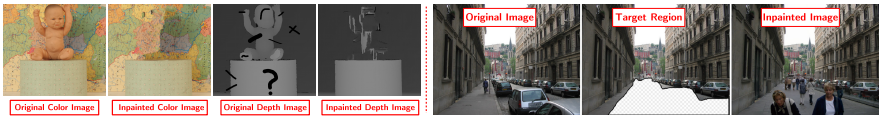

Fig. 1.3: **Left:** results of exemplar-based inpainting [36] applied to RGB and depth images. Note that the objective is to remove the object (baby) from both the RGB and depth images and to fill the already existing holes (pre-removal) in the depth image. The approach is significantly more effective when applied to colour images. **Right:** result of exemplar-based inpainting using an external database [60].

composed of two separate networks, a content encoder, based on [113], which is tasked to preserve contextual structures within the image, and a texture network, which enforces similarity of the fine texture within and without the target region using neural patches [95]. The model is capable of completing higher resolution images than [113, 150] but at the cost of greater inference time since the final output is not achievable via a single forward pass through the network.

More recently, significantly better results have been achieved using [73], which improves on the model in [113] by introducing global and local discriminators as adversarial loss components. The global discriminator assesses whether the completed image is coherent as a whole, while the local discriminator concentrates on small areas within the target region to enforce local consistency. Similarly, [151] trains a fully convolutional neural network capable of not only synthesizing geometric image structures but also explicitly using image features surrounding the target region as reference during training to make better predictions.

While these learning approaches are highly capable of generating perceptually plausible outputs despite the significant corruption applied to the input, when it comes to depth, they are incapable of producing high quality outputs due in part to the significantly higher number of target regions (holes) both large and small over the smoother surfaces in depth images. Examples of these novel approaches applied to depth images can be seen in Figure 1.4, which indicates how ineffective learning-based RGB image inpainting approaches can be within the depth modality [4].

While RGB completion techniques in various forms have previously been used with or without modifications [100, 144, 154] to complete depth images, significant differences between RGB and depth images prevent a successful deployment of RGB inpainting techniques to perform depth completion. For instance, the lack of reflectance colour texture in depth images, large featureless regions within the depth, overly smooth or blurred depth which can obscure object geometry, holes overlapping with object boundaries and unclear stopping points that mark the termination of structure continuation all contribute to the fact that specifically-designed approaches are required to handle the completion of depth images, leading to the importance of the existing literature on depth completion.

Consequently, RGB inpainting is not the focus of this chapter and is only covered here to give context to the relevant literature on depth completion. As such, the reader is invited to refer to the wide-expanding surveys that already exist on the

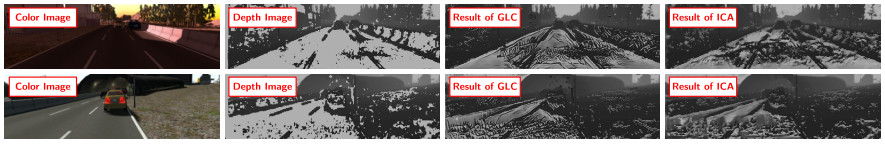

Fig. 1.4: Results of global and local completion (GLC) [73] compared to inpainting with contextual attention (ICA) [151]) applied to depth images.

issues of texture synthesis and inpainting within the context of RGB images [58, 78, 131, 139].

1.3.2 Depth Filling

One of the most important steps in addressing any problem, such as that of depth completion, is to focus on how the problem can be formulated. Numerous research works have attempted to solve the depth filling problem by concentrating on different challenges within the domain. In this section, a general overview of the most common formulations of the depth completion problem is presented before moving on to discussing a brief taxonomy of the depth filling literature.

1.3.2.1 Problem Formulation

Reformulating any ill-posed problem such as depth completion can lead to solutions suitable for particular requirements pertaining to certain situations, including time, computation, accuracy and alike. In this section, some of the most common ways in which depth filling has been posed and solved as a problem, and the effects each reformulation can have on the results are discussed.

Formulating the image completion and de-noising problem as **anisotropic diffusion** [115] has proven very successful in the context of RGB images [10, 15, 22]. Such solutions have therefore also made their way into the domain of depth image completion, since the smoothing and edge-preserving qualities of the diffusion-based solutions are highly desirable when dealing with depth information. This is primarily because image gradients are stronger where depth discontinuities are most likely and scene depth is often locally smooth within a single object.

Anisotropic diffusion is a non-linear partial differential equation scheme [115] which can be described as a space-variant transformation of an input image. It can therefore generate a family of smoothed parametrized images, each of which corresponds with a filter that depends on the local statistics of the input image.

More formally, if $I(\cdot,t)$ is a family of parametrized images, then the anisotropic diffusion is:

$$I_t = div(c(x,y,t)\nabla I) = c(x,y,t)\Delta I = \nabla c \cdot \nabla I, \tag{1.1}$$

where div is the divergence operator, ∇ and Δ denote the gradient and Laplacian operators respectively, and $c(x,y,t)$ is the diffusion coefficient, which can be a constant or a function of the image gradient.

In [136], Eqn. 1.1 is discretized via a 4-neighbourhood scheme, and the corresponding RGB image is used to guide the depth diffusion in an iterative process. The depth image is completed at a lower spatial resolution, and the iterative colour-guided anisotropic diffusion subsequently corrects the depth image as it is upsampled step by step.

The work of [107] demonstrates another example of the use of diffusion in depth completion. The process begins by extracting edges from the corresponding RGB image captured via a structured-light device, and then the smooth and edge regions undergo different diffusion algorithms. The separation of these regions before the diffusion process is performed based on the observation that surfaces which need to be smooth in the depth may be textured in the RGB image, and object boundaries within the depth image can be missed during the RGB edge extraction process due to the potentially low contrast in the RGB view of the scene.

While smooth surfaces and strong object boundaries can be very desirable traits in a depth image, the implementation of an anisotropic diffusion method requires discretization, which can lead to numerical stability issues and is computationally intensive. The longer run-time of diffusion-based methods make them intractable within real-time applications.

Energy minimization is another formulation of the completion problem which has seen significant success in the domain of RGB image inpainting [12, 140] and has consequently been used in depth filling as well.

Energy minimization relies on certain assumptions made about the image, using which an energy function is designed. Essentially, prior knowledge about images and sensing devices is modelled via regularization terms that form the energy function. This function is subsequently optimized, which leads to the completion and enhancement of the image based on the criteria set by the different terms within the function. The approaches addressing the depth completion problem in this manner often produce accurate and plausible results but more importantly, the capability of these approaches to focus on specific features within the image based on the terms added to the energy function is highly advantageous.

For example, the energy function in [31] models the common features of a depth image captured using a structured light device. The noise model of the device and the structure information of the depth image are taken into account using terms added to the energy function, performing regularization during the minimization process. Similarly, [103] assumes a linear correlation between depth and RGB values within small local neighbourhoods. An additional regularization term based on [11] enforces sparsity in vertical and horizontal gradients of the depth image, resulting in sharper object boundaries with less noise. The energy function in [63] includes a data term that favours pixels surrounding hole boundaries and a smoothing term that encourages locally smoother surfaces within the depth image. While

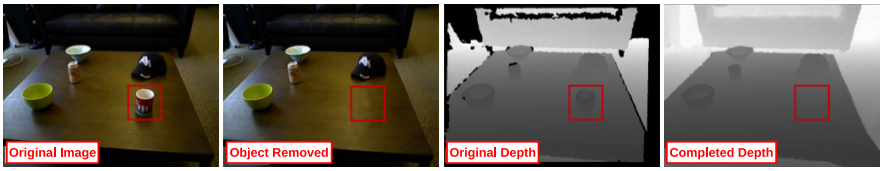

Fig. 1.5: Example of the results of [7]. An object has been removed from the RGB-D image and the missing values in the depth image have been completed.

this leads to better geometric and structural coherency within the scene, surface relief and texture is lost in the resulting depth image.

The lack of accurate surface relief and texture is in fact a very common challenge with many depth completion techniques. This issue can be addressed by solving depth completion as an **exemplar-based inpainting** problem, which has seen enormous success in RGB images [36]. Most exemplar-based inpainting techniques operate on the assumption that the information needed to complete the target region (with respect to both texture and structural continuity) is contained within the known regions of the image. As a result, plausible image completion can be achieved, at least in part, by copying and pasting patches, sometimes in a very specific order [36], from the known regions of the image into the target region.

However, there can be major pitfalls with using an exemplar-based technique to complete missing values in a depth image. For instance, the lack of reflectance colour texture on a smooth surface which leads to unified depth can confuse an exemplar-based approach to a great degree. As seen in Figure 1.3 (left), the notable exemplar-based inpainting method of [36] is capable of filling the target region post object removal from the RGB image in a plausible way due to existence of visible colour texture in the background but for a depth image, where no colour texture is present and the background only consists of a flat plane, the results are not nearly as impressive (Figure 1.3 - left). Please note that the goal is to remove an object (the baby) from both the RGB and depth images and plausibly complete the remaining holes post removal and at the same time fill the existing holes in the depth image (represented by black markings on the depth image).

Nevertheless, just as various depth completion techniques take advantage of other inpainting approaches such as [133], with or without modifications [100, 144, 154], exemplar-based image inpainting has also left its mark on depth completion.

For instance, in [7], object removal and depth completion of RGB-D images is carried out by decomposing the image into separate high and low spatial frequency components by means of butterworth filtering in Fourier space. After the disentanglement of high and low frequency images, the high frequency information (object boundaries and texture relief) is filled using a classic texture synthesis method [43] reformulated as a pixel-by-pixel exemplar-based inpainting approach and enhanced by means of query expansion within the search space, and the low frequency component (underlying shape geometry) is completed via [2]. The results are then re-

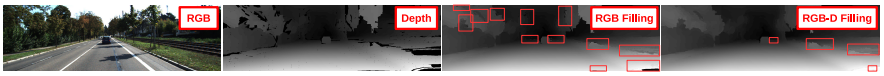

Fig. 1.6: Example of the results of exemplar-based RGB-D completion [5] as opposed to exemplar-based RGB completion applied to depth images from the Middlebury dataset [66]. The artefacts are marked with red boxes.

combined in the frequency domain to generate the final output. As seen in Figure 1.5, the produced images are sharp and with no additional artefacts.

Exemplar-based completion also makes an appearance in [9], which performs object removal in multi-view images with an extracted depth image, and uses both structure propagation and structure-guided filling to complete the images. The target region is completed in one of a set of multi-view photographs casually taken in a scene. The obtained images are first used to estimate depth via structure from motion. Structure propagation and structure-guided completion are employed to create the final results after an initial RGB-D completion step. The individual steps of this algorithm use the inpainting method in [140], and the patch-based exemplar-based completion approach of [38] to generate the results.

The work in [5], extends on the seminal RGB inpainting technique of [36] to create an exemplar-based approach explicitly designed to complete depth images. This is achieved by adding specific terms focusing on the characteristics of depth images into the priority function, which determines which patches take precedence in the filling order. By introducing texture and boundary terms, [5] ensures that surface relief and texture is well preserved in the depth image after completion, leading to more plausible results with fewer artefacts. As seen in Figure 1.6, the RGB completion technique [36] applied to depth images produces many undesirable artefacts while [5] generates sharper depth outputs.

Even though solving the depth filling problem using an exemplar-based framework has the potential to produce outputs in which structural continuity within the scene is preserved and granular relief texture is accurately and consistently replicated in the missing depth regions, there are still many challenges the completion process must contend with. For instance, if the scene depth is not of a fronto-parallel view, there is no guarantee that correct depth values can be predicted for the missing regions via patch sampling even if the patches undergo different transformations such as rotation, scale, shear, aspect ratio, keystone corrections, gain and bias colour adjustments, and other photometric transformations in the search space when trying to find similar patches to sample from [4].

To combat some of these issues, **matrix completion** has recently emerged as an interesting formulation of the image completion problem, especially since it has been observed [104] that similar patches in an RGB-D image lie in a low-dimensional subspace and can be approximated by a matrix with a low rank. The approach in [104] presents a linear algebraic method for low-rank matrix completion-based depth image enhancement to simultaneously remove noise and complete depth images using the corresponding RGB images, even if they contain heavily

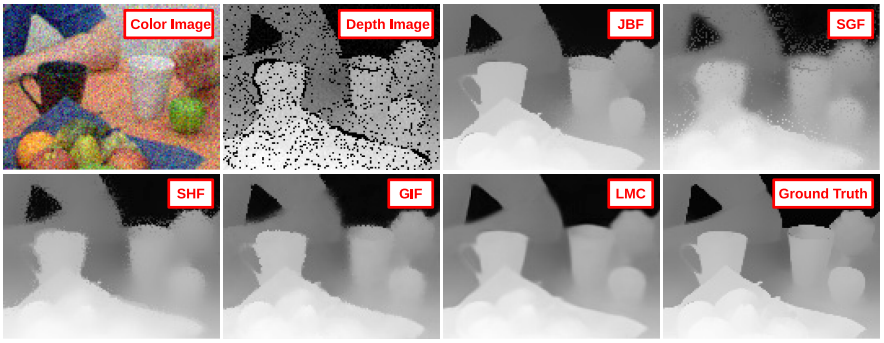

Fig. 1.7: Demonstrating the results of the matrix completion technique of [104] using low-rank operations (denoted by LMC) compared to joint bilateral filtering (JBF) [122], structure guided fusion (SGF) [120], spatio-temporal hole filling (SHF) [25] and guided inpainting and filtering (GIF) [100].

visible noise. In order to accomplish simultaneous de-noising and completion, the low-rank subspace constraint is enforced on a matrix with RGB-D patches via incomplete factorization, which results in capturing the potentially scene-dependent image structures both in the depth and colour space.

The rank differs from patch to patch depending on the image structures, so a method is proposed to automatically estimate a rank number based on the data. Figure 1.7 demonstrates the performance capabilities of this approach compared to other depth completion methods, such as joint bilateral filtering (JBF) [122], structure guided fusion (SGF) [120], spatio-temporal hole filling (SHF) [25] and guided inpainting and filtering (GIF) [100]. This approach [104] generates particularly impressive results in that the input RGB image is very noisy (Figure 1.7 - Color Image). Before the comparisons, a de-noising method [37] is applied to the noisy RGB image used as an input for the comparators.

The work in [145] points out, however, that the low-rank assumption does not fully take advantage of the characteristics of depth images. Sparse gradient regularization can naively penalize non-zero gradients within the image but based on statistical observations, it is demonstrated that despite most pixels having zero gradients, there is still a relatively significant number of pixels with gradients of 1. Therefore, a low-gradient regularization scheme is proposed in which the penalty for gradient 1 is reduced while non-zero gradients are penalized to allow for gradual changes within the depth image. This regularization approach is subsequently integrated with the low-rank regularization for depth completion.

More recently, with the advent of deep neural network, many image generation problems such as RGB inpainting [73, 113, 146, 150, 151] are essentially formulated as an **image-to-image translation** problem using a mapping function approximated by a deep network directly supervised on ground truth samples. However, as seen in Figure 1.4, networks designed to complete RGB images might not work well

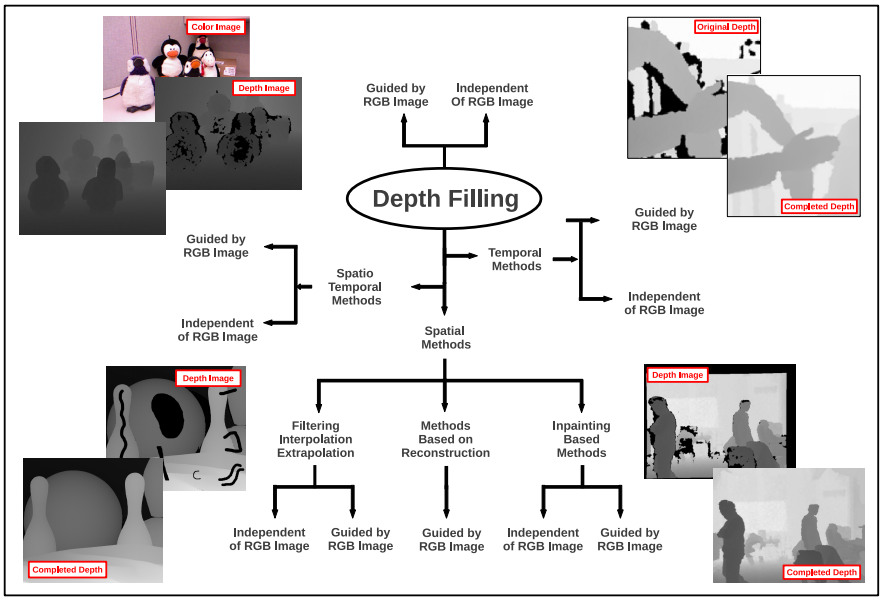

Fig. 1.8: A diagrammatic taxonomy of depth filling based on inputs and the information domain used during the completion process.

when it comes to depth. A significant obstacle to creating a neural network trained to complete scene depth is the lack of hole-free ground truth depth images available.

To overcome this problem, [157] creates a dataset of RGB-D images based on available surface meshes reconstructed from multi-view RGB-D scans of large environments [29]. Reconstructed meshes from different camera poses are rendered, which produces a supply of complete RGB-D images. This data is subsequently utilized to train a network that produces dense surface normals and occlusion boundaries. The outputs are then combined with raw depth data provided by a consumer RGB-D sensor to predict all depth pixels including those missing (holes).

While the formulation of a problem plays a significant role in the quality of the solution, the desired outcome of depth completion is highly dependent on a variety of factors, including the availability of the input data, the information domain, computational requirements and alike. In the following section, a brief discussion of the most successful depth completion techniques in the literature is provided.

1.3.2.2 A Taxonomy of Depth Completion

Within the literature, different depth completion techniques are often designed around the information domain available as the input or required as the output. Some techniques only utilize the spatial information locally contained within the image, while some take advantage of the temporal information extracted from a video se-

Categories	Subcategories	Examples
Spatial based methods	Filtering, Interpolation, Extrapolation	[3, 89, 92, 108, 117, 148]
	Inpainting based	[39, 64, 100, 120, 136]
	Reconstruction based	[30, 31, 103, 147]
Temporal based methods		[13, 48, 75, 106, 130]
Spatio-temporal based methods		[24, 25, 122, 137]

Table 1.1: A taxonomy of depth filling completion based on the information domain used during the filling process.

quence used to complete or homogenize the scene depth, and there are some that are based on a combination of both (Figure 1.8 and Table 1.1).

Spatial-based depth completion approaches use the neighbouring pixel values and other information available in a single RGB-D image to complete any missing or invalid data in the depth image. Even though there are clear limitations to using this type of approach, such as a possible lack of specific information that can be construed as useful to a particular target region (hole) in the scene depth, there are many important advantages. For instance, when temporal and motion information is taken into consideration for depth completion, filling one frame in a video requires processing multiple consecutive frames around it and so either the processing has to be done off-line or if real-time results are needed, the results of each frame will appear with a delay. However, if there is no dependence on other frames, with an efficient spatial-based method, real-time results can be generated without any delay.

One of the simplest, yet not always the best, approaches to using the spatial information within a single RGB-D frame is to employ a *filtering* mechanism to scene depth. Some common filters of choice would be the median filter [88] or the Gaussian filter [155] but with their use comes significant blurring effects and loss of texture and sharp object boundaries. However, there are image filtering techniques with edge-preserving qualities, such as the bilateral filter [135] and non-local filter [21]. On the other hand, these filters will not only preserve edges at object boundaries but the undesirable depth discontinuities caused by depth sensing issues as well.

There have been attempts to use the visual information present in the colour component of the RGB-D image to improve the accuracy of the depth completion results within or near object boundaries. This notion has also been utilized to reduce the noise in depth images generated by upsampling procedures [49, 82], where the goal is to increase the sharpness, accuracy, and the resolution of the depth image. Moreover, it can also be used to assist filtering approaches, as seen in methods such as joint-bilateral filtering [116], joint trilateral filtering [102] and alike.

A fast and non-approximate linear-time guided filtering method is proposed in [61]. The output is generated based on the contents of a guidance image. It can transfer the structures of the guidance image into the output and has edge-preserving qualities like the bilateral filter but can perform even better near object boundaries

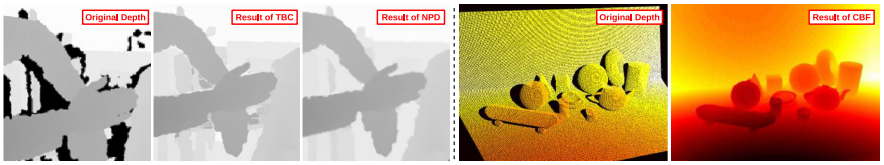

Fig. 1.9: **Left:** example of the result of Neighbouring Pixel Distribution (NPD) approach [148] compared to Temporal Based Completion (TBC) of [106] ;**Right:** example of depth completion using cross bilateral filtering [112].

and edges by avoiding reversal artefacts. Due to its efficiency and performance, it has been used as the basis for several depth completion methods [100, 147].

The approach in [148] completes depth images based on the depth distribution of pixels adjacent to the holes after labelling each hole and dilating each labelled hole to get the value of the surrounding pixels. Cross-bilateral filtering is subsequently used to refine the results. In Figure 1.9 (left), the results are compared with the temporal based method in [106], which will be discussed subsequently.

Similarly, in [92], object boundaries are first extracted, and then a discontinuity-adaptive smoothing filter is applied based on the distance of the object boundary and the quantity of depth discontinuities. The approach in [112] proposes a propagation method, inspired by [110], that makes use of a cross bilateral filter to fill the holes in the image (as seen in Figure 1.9 - right).

In [108], an approach based on weighted mode filtering and a joint histogram of the RGB and depth image is used. A weight value is calculated based on the colour similarity between the target and neighbouring pixels on the RGB image and used for counting each bin on the joint histogram of the depth image. [109], on the other hand, uses adaptive cross-trilateral median filtering to reduce the noise and inaccuracies commonly found in scene depth obtained via stereo correspondence. Parameters of the filter are adapted to the local structures, and a confidence kernel is employed in selecting the filter weights to reduce the number of mismatches.

In an attempt to handle the false contours and noisy artefacts in depth estimated via stereo correspondence, [89] employs a joint multilateral filter that consists of kernels measuring proximity of depth samples, similarity between the sample values, and similarity between the corresponding colour values. The shape of the filter is adaptive to brightness variations.

Various *interpolation* and *extrapolation* methods using the spatial information within RGB-D images have also appeared in the depth completion literature. For instance, an object-aware non-parametric interpolation method is proposed in [3], which utilizes a segmentation step [8] and redefines and identifies holes within a set of 12 completion cases with each hole existing in a single row of a single object. The depth pattern is then propagated into hole regions accordingly. Figure 1.10 demonstrates the efficacy of the approach [3] compared to [2, 7, 63, 100, 133]. Additionally, the approach [3] functions in a manner of milliseconds, making it highly effective in real-time application, as seen in Table 1.2.

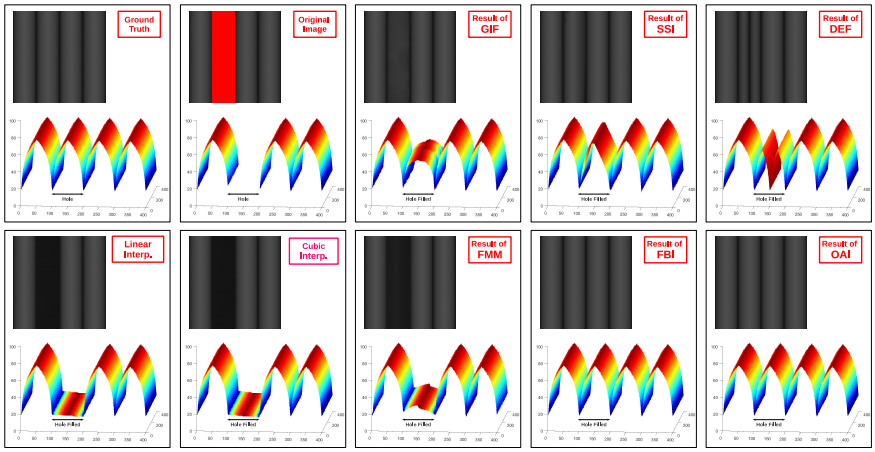

Fig. 1.10: Comparing the results of guided inpainting and filtering (GIF) [100], second-order smoothing inpainting (SSI) [63], fast marching based inpainting (FMM) [133], Fourier-based inpainting (FBI) [7], diffusion-based exemplar filling (DEF) [2], object-aware interpolation (OAI) [3] and linear and cubic interpolation using a synthetic test image with available ground truth depth.

There are other interpolation techniques that complete depth images horizontally or vertically within target boundaries by calculating a normalized distance between opposite points of the border (horizontally or vertically) and interpolating the pixels accordingly [117]. These approaches can face performance challenges when the target region (hole) covers parts of certain structures that are neither horizontal nor vertical. To prevent this potential issue, [117] proposes a multi-directional extrapolation technique that uses the neighbouring texture features to estimate the direction in which extrapolation is to take place, rather than using the classic horizontal or vertical directions that create obvious deficiencies in the completed image.

Similarly, [51] presents a segmentation-based interpolation technique to upsample, refine, and enhance depth images. The strategy uses segmentation methods that combine depth and RGB information [35, 105] in the presence of texture. Alternatively, when the image is not highly-textured, segmentation techniques based on graph cuts [47] can be used to identify the surfaces and objects in the RGB im-

Method	Error (lower, better)		Run-time (*ms*)	Method	Error (lower, better)		Run-time (*ms*)
	RMSE	PBMP			RMSE	PBMP	
Linear Inter.	1.3082	0.0246	25.12	Cubic Inter.	1.3501	0.0236	27.85
GIF [100]	0.7797	0.0383	3.521e3	SSI [63]	3.7382	0.0245	51.56e3
FMM [133]	1.0117	0.0365	4.31e3	DEF [2]	0.6188	0.0030	8.25e5
FBI [7]	0.6944	0.0058	3.84e6	OAI [3]	0.4869	0.0016	99.09

Table 1.2: Average RMSE, PBMP, & run-time (test images from Middlebury [66]).

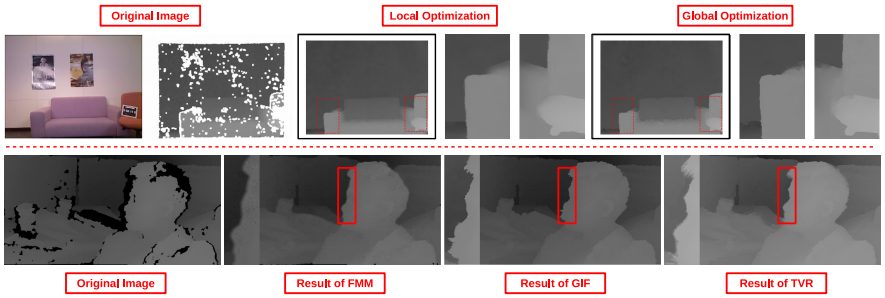

Fig. 1.11: **Top:** local and global framework of [31]. The energy function is made up of a fidelity term (generated depth data characteristics) and a regularization term (joint-bilateral and joint-trilateral kernels). Local filtering can be used instead of global filtering to make parallelization possible. **Bottom:** example of the results of depth completion using energy minimization with TV regularization (TVR) [103] compared to fast marching method based inpainting (FMM) [133] and guided inpainting and filtering (GIF) [100]. The energy function assumes that in small local neighbourhoods, depth and colour values are linearly correlated.

age, which are assumed to align with those in the depth image. The low-resolution depth image is later projected on the segmented RGB image and interpolation is subsequently performed on the output.

While spatial-based depth completion strategies using filtering, interpolation, and extrapolation techniques are among the most used and most efficient methods, traditional *inpainting-based* techniques (normally used for RGB images, Section 1.3.1) can yield more promising results in terms of accuracy and plausibility despite being computationally expensive.

The approach in [120] attempts to recover the missing depth information using a fusion-based method integrated with a non-local filtering strategy. Object boundaries and other stopping points that mark the termination of structure continuation process are not easy to locate in depth images which generally have little or no texture, or the boundaries or stopping points might be in the target region within the depth image. The RGB image is thus used to assist with spotting the boundaries, and their corresponding positions in the depth image are estimated according to calibration parameters. The inpainting framework follows the work of [22] that takes advantage of a scheme similar to the non-local means scheme to make more accurate predictions for pixel values based on image textures. To solve the issue of structure propagation termination, a weight function is proposed in the inpainting framework that takes the geometric distance, depth similarity, and structure information within the RGB image into account.

The fast marching method-based inpainting of [133] has achieved promising success in RGB inpainting (Section 1.3.1). The work of [100] improves upon this approach for depth completion by using the RGB image to guide the depth inpainting process. By assuming that the adjacent pixels that have similar colour values have

a higher probability of having similar depth values as well, an additional *colour term* is introduced into the weighting function to increase the contribution of the pixels with the same colour. The order of filling is also changed so that the pixels near edges and object boundaries are filled later, in order to produce sharper edges. However, even with all the improvements, this guided depth inpainting method is still not immune to noise and added artefacts around object boundaries (as seen in Figures 1.11 - bottom, 1.7 and Figure 1.12); therefore, the guided filter [61] is used in the post-processing stage to refine the depth image.

The work in [144] introduces an exemplar-based inpainting method to prevent the common blurring effects produced while completing the scene depth in novel views synthesized through depth image-based rendering. In the two separate stages of warped depth image hole filling and warped RGB image completion, the focus is mainly on depth-assisted colour completion with texture. The depth image is assumed to be only a gray-scale image with no texture, and is therefore filled using any available background information (i.e., depth pixels are filled by being assigned the minimum of the neighbouring values). The assumptions that depth images have no texture, that texture and relief are not of any significant importance in depth images, and depth holes can be plausibly filled using neighbouring background values are obviously not true, and lead to ignoring the utter importance of accurate 3D information in the state of the art. As a result, although the inpainting method proposed in [144] to complete newly synthesized views based on depth is reasonable, the depth filling itself is lacking.

An anisotropic diffusion-based method is proposed in [136] that can have real-time capabilities by means of a GPU. The RGB image is used to guide the diffusion in the depth image, which saves computation in the multi-scale pyramid scheme since the RGB image does not change. In order to guarantee the alignment of the object boundaries in the RGB and the depth image, anisotropic diffusion is also applied to object boundaries.

Although inpainting based depth filling techniques can produce reasonable and efficient results, there is a possibility of blurring, ringing, and added artefacts especially around object boundaries, sharp discontinuities and highly textured regions. In *reconstruction-based* methods, however, missing depth values are predicted using common synthesis approaches. Since a closed-loop strategy is mostly used to resolve the reconstruction coefficients in terms of the minimization of residuals, higher levels of accuracy can be accomplished in depth completion. There are numerous different models found in the literature that are used to represent the depth completion problem as such.

For instance, in [30, 31], energy minimization is used to solve the depth completion problem, specifically depth generated by consumer depth sensors. The energy function consists of a fidelity term that considers the characteristics of consumer device generated depth data and a regularization term that incorporates the joint-bilateral kernel and the joint-trilateral kernel. The joint-bilateral filter is tuned to incorporate the structure information and the joint-trilateral kernel is adapted to the noise model of consumer device generated depth data. Since the approach is relatively computationally-expensive, local filtering is used to approximate the global

optimization framework in order to make parallelization possible, which brings forth the long-pondered question of accuracy versus efficiency. A comparison between examples of the results generated through both local and global frameworks is seen in Figure 1.11 (top).

The work of [93] in image matting inspired [103] to design an energy function based on the assumption that in small local neighbourhoods, there is a linear correlation between depth and RGB values. To remove noise and create sharper object boundaries and edges, a regularization term originally proposed in [11] is added to the energy function, which makes the gradient of the depth image both horizontally and vertically sparse, resulting in less noise and sharper edges. A comparison between the results of this method and inpainting methods in [133] and [100] is shown in Figure 1.11 (bottom).

Figure 1.12 contains a qualitative comparison of some of the spatial-based depth filling methods [3, 3, 63, 100], RGB completion techniques [2, 36, 133], and bilinear interpolation over examples from the Middlebury dataset [66]. Table 1.2 presents the numerical evaluation of the same approaches by comparing their Root Mean Square Error (RMSE), Percentage of Bad Matching Pixels (PBMP), and their run-time. As you can see, even though spatial-based methods are certainly capable of achieving real-time results (unlike temporal-based methods), the current literature epitomizes the long-standing trade-off between accuracy and efficiency. Many of these methods are capable of filling only small holes [3] and others are extremely inefficient [7]. Any future work will need to work towards achieving higher standards of accuracy and plausibility in shorter periods of time.

Certain depth completion techniques in the literature take advantage of the motion and temporal information contained within a video sequence to complete and refine depth images [13, 106]. One of these **temporal-based** approaches, commonly used as a comparator in the literature, is the method proposed in [106] which utilizes motion information and the difference between the depth values in the current image and those in the consecutive frames to fill holes by giving the pixels the weighted average values of the corresponding pixels in other frames. Although the results are mostly plausible, one drawback is that the value of the edges of objects cannot be accurately estimated to an acceptable level (Figures 1.9 - left), other than the fact that there is a need for a sequence of depth images, and therefore, the holes in a single depth image cannot be filled. Moreover, when the colour information does not correspond with the depth data, the results often contain invalid depth values.

The well-known KinectFusion approach of [75] takes advantage of the neighbouring frames to complete the missing depth during real-time 3D reconstruction. However, camera motion and a static scene are of utmost importance and despite being robust, the approach cannot be utilized for a static view of a scene without any camera motion. In [13], missing depth regions are grouped into one of two categories: the ones created as a result of occlusion by foreground objects, assumed to be in motion, and the holes created by reflective surfaces and other random factors. Subsequently, they use the deepest neighbouring values to fill pixels according to the groups they are placed in. Even though the assumptions might be true in many real-

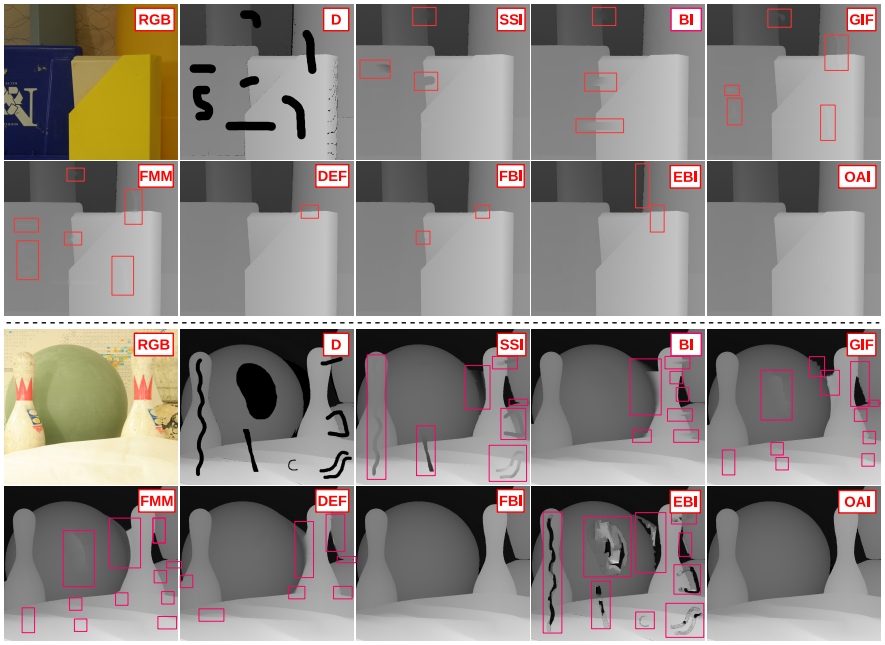

Fig. 1.12: Comparing the results of guided inpainting and filtering (GIF) [100], second-order smoothing inpainting (SSI) [63], fast marching based inpainting (FMM) [133], Fourier-based inpainting (FBI) [7], diffusion-based exemplar filling (DEF) [2], object-aware interpolation (OAI) [3] and bilinear interpolation (BI) using examples from the Middlebury dataset [66].

life scenarios, they are not universal, and static objects can be the cause of missing or invalid data in depth images captured via many consumer depth sensors.

The approach in [48] focuses on repairing the inconsistencies in depth videos. Depth values of certain objects in one frame sometimes vary from the values of the same objects in a neighbouring frame, while the planar existence of the object has not changed. An adaptive temporal filtering is thus proposed based on the correspondence between depth and RGB sequences. [130] notes that the challenge in detecting and mending temporal inconsistencies in depth videos is due to the dynamic content and outliers. Consequently, they propose using the intrinsic static structure, which is initialized by taking the first frame and refined as more frames become available. The depth values are then enhanced by combining the input depth and the intrinsic static structure, the weight of which depends on the probability of the input value belonging to the structure. As seen in Figure 1.13 (left), the method proposed in [130] does not introduce artefacts into the results due to motion delay because temporal consistency is only enforced on static regions, as opposed to [48], which applies temporal filtering to all regions.

Fig. 1.13: **Left:** example of the results of completion based on intrinsic static structure (ISS) [130] compared to adaptive temporal filtering (ATF) [48]; **Right:** example of the results of spatio-temporal completion (STC) [25] compared to exemplar based inpainting (EBI) [36] on still frames.

Temporal-based methods generate reasonable results even when spatial-based approaches are unable to, and are necessary when depth consistency and homogeneity is important in a depth sequence, which it often is. On the other hand, the dependency on other frames is a hindrance that causes delays or renders the method only applicable as an off-line approach. Moreover, there are many scenarios where a depth sequence is simply not available but a single depth image still needs to be completed. **Spatio-temporal** completion approaches, however, combine the elements of the spatial and temporal based methods to fill holes in depth images [25, 137].

In [137], the process of depth completion is carried out in two stages. First, a *deepest depth image* is generated by combining the spatio-temporal information in the depth and RGB images and used to fill the holes. Subsequently, the filled depth image is enhanced based on the joint information of geometry and colour. To preserve local features of the depth image, filters adapted to RGB image features are utilized. In another widely-used method, [24] use an adaptive spatio-temporal approach to fill depth holes utilizing bilateral and Kalman filters. The approach is made up of three blocks: an adaptive joint bilateral filter that combines the depth and colour information is used, random fluctuations of pixel values are subsequently handled by applying an adaptive Kalman filter on each pixel, and finally, an interpolation system uses the stable values in the regions neighbouring the holes provided by the previous blocks, and by means of a 2D Gaussian kernel, fills the missing depth values.

In another method [25], scene depth is completed using a joint-bilateral filter applied to neighbouring pixels, the weights of which are determined based on visual data, depth information, and a temporal consistency map that is created to track the reliability of the depth values near the hole regions. The resulting values are taken into account when filtering successive frames, and iterative filtering can ensure increasing accuracy as new samples are acquired and filtered. As seen in Figure 1.13 (right), the results are superior to the ones produced by the inpainting algorithm proposed in [36].

Improvements made to what can be obtained from a regular video camera alongside a time-of-flight camera is discussed in [122], and the main focus of the work is on depth upsampling and colour/depth alignment. However, one of the issues addressed is depth completion, which is performed via a multi-scale technique following the works in [57] and [84]. The output undergoes joint bilateral filtering

and spatio-temporal processing to remove noise by averaging values from several consecutive frames.

The approach presented in [74] uses a sequence of frames to locate outliers with respect to depth consistency within the frame, and utilizes an improved and more efficient regression technique using least median of squares (LMedS) [124] to fill holes and replace outliers with valid depth values. The approach is capable of hole filling and sharp depth refinement within a sequence of frames but can fail in the presence of invalid depth shared between frames or sudden changes in depth due to fast moving dynamic objects within the scene.

While depth completion can be a useful process for creating full dense depth for various vision-based application, learning-based monocular depth estimation techniques can be an invaluable tool that can provide hole-free scene depth in a cheap and efficient manner, completely removing the need for any depth completion in the process. In the next section, a brief outline of the advances made in the field of monocular depth estimation is presented.

1.4 Monocular Depth Estimation

Over the past few years, research into monocular depth estimation, i.e. predicting complete scene depth from a single RGB image, has significantly escalated [44, 50, 55, 87, 99, 143]. Using off-line model training based on ground truth depth data, monocular depth prediction has been made possible [44, 45, 87, 99, 162] sometimes with results surpassing those of more classical depth estimation techniques. Ground truth depth, however, is extremely difficult and expensive to acquire, and when it is obtained it is often sparse and flawed, constraining the practical use of monocular depth estimation in real-world applications. Solutions to this problem of data scarcity include the possibility of using synthetic data containing sharp pixel-perfect scene depth [6] for training or completely dispensing with using ground truth depth, and instead utilizing a secondary supervisory signal during training which indirectly results in producing the desired depth [32, 50, 55, 143].

In the following, a brief description of monocular depth estimation techniques within three relevant areas is provided: approaches utilizing hand-crafted features based on monocular cues within the RGB image, approaches based on graphical models and finally techniques using deep neural networks trained in various ways to estimate depth from a single image.

1.4.1 Hand-Crafted Features

While binocular vision is commonly associated with depth perception in humans and machines, estimating depth from a single image based on monocular cues and features is technically possible for both humans and machines, even if the results

are not very accurate. Such monocular cues include size considering visual angles, grain, and motion parallax. Monocular depth estimation techniques have utilized such features to estimate depth from a single RGB image.

Based on the assumption that the geometric information contained within a scene combined with motion extracted from a sequence can be valuable features for 3D reconstruction, [70] estimates depth based on temporal continuity and geometric perspective. In [153], different cues such as motion, colour and contrast are combined to extract the foreground layer, which is then used to estimate depth. Motion parameters and optical flow are calculated using structure from motion.

In [67, 68], an assumption of ground-vertical geometric structure is used as the basis to construct a basic 3D model from a single photograph. This is accomplished by labelling the image according to pre-defined geometric classes and subsequently creating a statistical model based on scene orientation. [81] proposes a non parametric approach based on SIFT Flow, where scene depth is reconstructed from an input RGB image by transferring the depth of multiple similar images and then applying warping and optimizing procedures. The work in [97] investigates using semantic scene segmentation results to guide the depth reconstruction process instead of directly predicting depth based on features present in the scene. The work in [87] also takes advantage of combining semantic object labels with depth features to aid in the depth estimation process.

It is important to note that predicting depth based on monocular cues within the scene is not robust enough to deal with complex and cluttered scenes even though approaches using such features have managed to produce promising results when it comes to scenes that contain clear pre-defined features and adhere to simple structural assumptions.

1.4.2 Graphical Models

Within the current literature on monocular depth estimation, there are approaches that take advantage of graphical models to recover scene depth. For instance, [40] introduces a dynamic Bayesian network model capable of reconstructing a 3D scene from a monocular image based on the assumption that all scenes contain a *floor-wall* geometry. The model distinguishes said floor-wall boundaries in each column of the image and using perspective geometry reconstructs a 3D representation of the scene. While the approach produces very promising results, the underlying assumption it is built on (indoor scenes framed by a floor-wall constraint) limits the capabilities of the approach.

The work in [127] utilizes a discriminatively-trained Markov Random Field (MRF) and linear regression to estimate depth. The images are segmented into homogeneous regions and the produced patches are used as super-pixels instead of pixels during the depth estimation process. This extended version of the approach [128] utilizes the MRF in order to combine planes predicted by the linear model to describe the 3D position and orientation of segmented patches within RGB images.

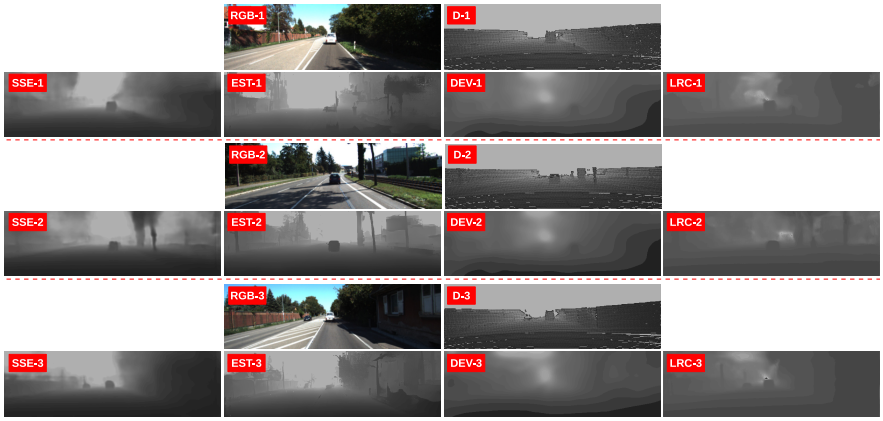

Fig. 1.14: Qualitative comparison of depth and ego-motion from video (DEV) [160], estimation based on left/right consistency (LRC) [55]; **SSE:** semi-supervised depth estimation (SSE) [86], depth estimation via style transfer (EST) [6].

Since depth is predicted locally, the combined output lacks global coherence. Additionally, the model is manually tuned which is a detriment against achieving a learning based system.

The method proposed in [62] presents cascaded classification models. The approach combines the tasks of scene categorization, object detection, multi-class image segmentation and, most relevant here, 3D reconstruction by coupling repeated instantiations of the sophisticated off-the-shelf classifiers in order to improve the overall performance at each level.

In [101], monocular depth estimation is formulated as an inference problem in a discrete/continuous Conditional Random Field (CRF) model, in which continuous variables encode the depth information associated with super-pixels from the input RGB image, and the discrete ones represent the relationships between the neighbouring super-pixels. Using input images with available ground truth depth, the unary potentials are calculated within a graphical model, in which the discrete/continuous optimization problem is solved with the aid of particle belief propagation [71, 114].

To better exploit the global structure of the scene, [162] proposes a hierarchical representation of the scene based on a CRF, which is capable of modelling local depth information along with mid-level and global scene structures. Not unlike [101], the model attempts to solve monocular depth estimation as an inference problem in a graphical model in which the edges provide an encoding of the interactions within and across the different layers of the proposed scene hierarchy.

More recently, [142] attempts to perform monocular depth estimation using sparse manual labels for object sizes within a given scene. Utilizing these manually estimated object sizes and the geometric relationship between them, a coarse depth image is primarily created. This depth output is subsequently refined using a

CRF that propagates the estimated depth values to generate the final depth image for the scene.

Monocular depth estimation techniques based on graphical models can produce impressive results but despite their excellent generalization capabilities, deep neural networks generate sharper and more accurate depth images, even though they can be prone to over-fitting and require larger quantities of training data.

1.4.3 Deep Neural Networks

Recent monocular depth estimation techniques using deep convolutional neural networks *directly supervised* using data with ground truth depth images have revolutionized the field by producing highly accurate results. For instance, the approach in [45] utilizes a multi-scale network that estimates a coarse global depth image and a second network that locally refines the depth image produced by the first network. The approach is extended in [44] to perform semantic segmentation and surface normal estimation as well as depth prediction.

In the work by [90], a fully-convolutional network is trained to estimate more accurate depth based on efficient feature up-sampling within the network architecture. In the up-sampling procedure, the outputs of four convolutional layers are fused by applying successive up-sampling operations. On the other hand, [98] points to the past successes that CRF-based methods have achieved in monocular depth estimation and presents a deep convolutional neural field model that takes advantage of the capabilities of a continuous CRF. The unary and pairwise potentials of the continuous CRF are learned in a deep network resulting in depth estimation for general scenes with no geometric priors.

The work in [26] trains a supervised model for estimation formulated as a pixel-wise classification task. This reformulation of the problem is made possible by transforming the continuous values in the ground truth depth images into class labels by discretizing the values into bins and labelling the bins based on their depth ranges. Solving depth estimation as a classification problem provides the possibility to obtain confidence values for predicted depth in the form of probability distributions. Using the obtained confidence values, an information gain loss is applied that enables selecting predictions that are close to ground-truth values during training.

Similarly, [94] also presents monocular depth estimation as a pixel-wise classification problem. Different side-outputs from the dilated convolutional neural network architecture are fused hierarchically to take advantage of multi-scale depth cues. Finally, soft-weighted-sum inference is used instead of the hard-max inference, which transforms the discretized depth score to continuous depth value. [69] attempts to solve the commonly-found issue of blurring effects in the results of most monocular depth estimation techniques by fusing features extracted at different scales from a network architecture that includes a multi-scale feature fusion module and a refinement module trained via an objective function that measures errors in depth, gradients and surface normals.

Method	Training Data	Error Metrics (lower, better)				Accuracy Metrics (higher, better)		
		Abs. Rel.	Sq. Rel.	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Data Set Mean [53]	[53]	0.403	0.530	8.709	0.403	0.593	0.776	0.878
Eigen et al. Coarse [44]	[53]	0.214	1.605	6.563	0.292	0.673	0.884	0.957
Eigen et al. Fine [44]	[53]	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Liu et al. [99]	[53]	0.202	1.614	6.523	0.275	0.678	0.895	0.965
Zhou et al. [160]	[53]	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Zhou et al. [160]	[53]+ [34]	0.198	1.836	6.565	0.275	0.718	0.901	0.960
Garg et al. [50]	[53]	0.152	1.226	5.849	0.246	0.784	0.921	0.967
Godard et al. [55]	[53]	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Godard et al. [55]	[53]+ [34]	0.124	1.076	5.311	0.219	0.847	0.942	0.973
Zhan et al. [152]	[53]	0.144	1.391	5.869	0.241	0.803	0.928	0.969
Atapour et al. [6]	S*	**0.110**	0.929	4.726	0.194	**0.923**	**0.967**	0.984
Kuznietsov et al. [86]	[53]	0.113	**0.741**	**4.621**	**0.189**	0.862	0.960	**0.986**

Table 1.3: Comparing the results of monocular depth estimation techniques over the KITTI dataset using the data split in [45]. S* denotes the synthetic data captured from a graphically rendered virtual environment.

While these approaches produce consistently more encouraging results than their predecessors, the main draw-back of any directly supervised depth estimation model is its dependence on large quantities of dense ground truth depth images for training. To combat this issue, synthetic depth images have recently received attention in the literature. [6] takes advantage of aligned nearly photo-realistic RGB images and their corresponding synthetic depth extracted from a graphically rendered virtual environment primarily designed for gaming for training a monocular depth estimation model. Additionally, a cycle-consistent adversarially trained style transfer approach [161] is used to deal with the domain shift between the synthetic images used for training and the real-world images the model is intended for in practice. Figure 1.14 (EST) contains examples of the results of this approach, which are very sharp and with clear object boundaries due to the fact that pixel-perfect synthetic depth has been used as training data. Likewise, [159] proposes a similar framework in which a separate network takes as its input both synthetic and real-world images and produces modified images which are then passed through a second network trained to perform monocular depth estimation.

While the use of synthetic training data can be a helpful solution to the issue of scarcity of ground truth depth, a new class of *indirectly supervised* monocular depth estimators have emerged that do not require ground truth depth, and calculate disparity by reconstructing the corresponding view within a stereo correspondence framework and thus use this view reconstruction as a secondary supervisory signal. For instance, the work in [143] proposes the Deep3D network, which learns to generate the right view from the left image used as the input, and in the process produces an intermediary disparity image. The model is trained on stereo pairs from a dataset of 3D movies to minimize the pixel-wise reconstruction loss of the generated right view compared to the ground truth right view. The desired output is a probabilistic disparity map that is used by a differentiable depth image-based rendering layer in the network architecture. While the results of the approach are very promising, the method is very memory intensive.

The approach in [50] follows a similar framework with a model very similar to an auto-encoder, in which the encoder is trained to estimate depth for the input image (left) by explicitly creating an inverse warp of the output image (right) in the decoder using the estimated depth and the known inter-view displacement, to reconstruct the input image. The technique uses an objective function similar to [143] but is not fully differentiable.

On the other hand, [55] argues that a simple image reconstruction as done in [50, 143] does not produce depth with high enough quality and uses bilinear sampling [77] and a left/right consistency check between the disparities produced relative to both the left and right images incorporated into training to produce better results. Examples of the results of this approach can be seen in Figure 1.14 (LRC). Even though the results are consistently impressive across different images, blurring effects within the depth image still persist.

In [152], the use of sequences of stereo image pairs is investigated for estimating depth and visual odometry. It is argued that utilizing stereo sequences as training data makes the model capable of considering both spatial (between left/right views) and temporal (forward/backward) warp error in its learning process, and can constrain scene depth and camera motion to remain within a reasonable scale.

While the approaches that benefit from view synthesis through learning the inter-view displacement and thus the disparity are capable of producing very accurate and consistent results and the required training data is abundant and easily obtainable, there are certain shortcomings. Firstly, the training data must consist of temporally aligned and rectified stereo images, and more importantly, in the presence of occluded regions (i.e. groups of pixels that are seen in one image but not the other), disparity calculations fail and meaningless values are generated (as seen in Figure 1.14 (LRC)).

On the other hand, the work in [160] estimates depth and camera motion from video by training depth and pose prediction networks, indirectly supervised via view synthesis. The results are favourable especially since they include ego-motion but the depth outputs are very blurry (as seen in Figure 1.14 (DEV)), do not consider occlusions and are dependent on camera parameters. The training in the work of [86] is supervised by sparse ground truth depth and the model is then enforced within a stereo framework via an image alignment loss to output dense depth. This enables the model to take advantage of both direct and indirect training, leading to higher fidelity depth outputs than most other comparators, as demonstrated in Figure 1.14 (SSE) and Table 1.3.

Within the literature, there are specific metrics that are commonly used to evaluate the performance of monocular depth estimation techniques. Given an estimated depth image d'_p and the corresponding ground truth depth d_p at pixel p with N being the total number of pixels for which valid ground truth and estimated depth exist, the following metrics are often used for performance evaluation in the literature:

- Absolute Relative Error (*Abs. Rel.*) [128]:

$$\frac{1}{N}\sum_{p}\frac{|d_p - d'_p|}{d_p} \tag{1.2}$$

- Squared Relative Error (*Sq. Rel.*) [128]:

$$\frac{1}{N}\sum_{p}\frac{||d_p - d'_p||^2}{d_p} \tag{1.3}$$

- Linear Root Mean Square Error (*RMSE*) [62]:

$$\sqrt{\frac{1}{N}\sum_{p}||d_p - d'_p||^2} \tag{1.4}$$

- Log Scale Invariant RMSE (*RMSE log*) [45]:

$$\sqrt{\frac{1}{N}\sum_{p}||log(d_p) - log(d'_p)||^2} \tag{1.5}$$

- Accuracy under a threshold [87]:

$$max(\frac{d'_p}{d_p}, \frac{d_p}{d'_p}) = \delta < threshold \tag{1.6}$$

Table 1.3 provides a quantitative analysis of the state-of-the-art approaches proposed in [6, 44, 50, 55, 86, 99, 152, 160]. The experiment is carried out on the test split used in [45], which has now become a convention for evaluations within the monocular depth estimation literature.

1.5 Conclusions

The primary focus of this chapter has been on techniques specifically designed to complete, enhance and refine depth images. This is particularly important as there are still several issues blocking the path to a perfect depth image such as missing data, invalid depth values, low resolution, and noise despite the significant efforts currently under way with regards to improving scene depth capture technologies.

The depth completion problem has been formulated in a variety of different ways, as has the related problem of RGB inpainting. Diffusion-based and energy minimization solutions to the problem are accurate with respect to structural continuity within the scene depth and can produce smooth surfaces within object boundaries, which can be a desirable trait for certain applications. However, these solutions are often inefficient, computationally expensive, and can bring forth implementation issues. Depth images can also be completed using an exemplar-based paradigm, which can accurately replicate object texture and relief as well as preserve the necessary geometric structures within the scene. There are, of course, a variety of other problem formulations, such as matrix completion, labelling, image-to-image mapping and alike, each focusing on certain traits within the desired scene depth.

Input requirements can also vary for different depth completion techniques. Depending on the acquisition method, depth is commonly obtained along with an aligned or easily alignable RGB image of the same scene. The information contained within this RGB image can be used to better guide the filling approach applied to the depth image. However, not all depth images are accompanied by a corresponding RGB image and processing the colour information can add to the computational requirements which may not be necessary depending on the application.

Within the depth completion literature, there are **spatial-based** methods that limit themselves to the information in the neighbouring regions adjacent to the holes in the depth image and possibly the accompanying RGB image. Some of these algorithms make use of *filtering* techniques, while some utilize *interpolation and extrapolation* approaches. The filtering, interpolation, and extrapolation methods can provide fast and clean results but suffer from issues like smoothed boundaries and blurred edges. Some research has been focused on using *inpainting-based* techniques, which have been proven successful in completing RGB images post object removal. Despite their satisfactory results, these methods are not all efficient and can generate additional artefacts near target and object boundaries. There are also *Reconstruction methods* that can generate accurate results using techniques inspired by scene synthesis methods. However, they are mostly difficult to implement and some have a strict dependency on the corresponding RGB view.

Temporal-based depth completion techniques make use of the motion information and the depth in the neighbouring frames of a video to fill the hole regions in the current depth frame. Sometimes the information in a single depth image is not enough to complete that image, which is where spatial-based methods fall short. Temporal-based approaches, however, do not suffer from this issue and have a larger supply of information at their disposal. This class of methods is still not perfect, and the need to process other frames to complete a depth image makes them more suited for off-line applications rather than real-time systems.

Additionally, various **spatio-temporal based** methods have been proposed that use both the spatial information contained within the scene depth and the temporal continuity extracted from a sequence to perform depth completion. Although these methods can be more accurate than spatial-based techniques and more efficient than temporal-based approaches, they still suffer from the issues of both these categories.

Furthermore, whilst future avenues of research need to explicitly consider computational efficiency, within the contemporary application domains of consumer depth cameras and stereo-based depth recovery, it is also highly likely they will be able to exploit temporal aspects of a live depth stream. It is thus possible that both temporal and spatio-temporal techniques will become the primary areas of growth within this domain over the coming years. This trend will be heavily supported by aspects of machine learning as innovative solutions to the issue of acquiring high-quality ground truth depth data become increasingly widespread.

Of course, another innovative solution to the problem of obtaining accurate 3D scenes is to provide a cheap and efficient alternative to the current 3D capture technologies that can produce high-fidelity hole-free scene depth, entirely circumnavigating the need for depth completion as a necessary post-processing operation.

Recent learning-based monocular depth estimation methods have made significant strides towards achieving this goal by providing accurate and plausible depth mostly in real time from a single RGB image.

References

1. Abrams, A., Hawley, C., Pless, R.: Heliometric stereo: Shape from sun position. Euro. Conf. Computer Vision pp. 357–370 (2012)
2. Arias, P., Facciolo, G., Caselles, V., Sapiro, G.: A variational framework for exemplar-based image inpainting. Computer Vision **93**(3), 319–347 (2011)
3. Atapour-Abarghouei, A., Breckon, T.: Depthcomp: Real-time depth image completion based on prior semantic scene segmentation. In: British Machine Vision Conference, pp. 1–12. BMVA (2017)
4. Atapour-Abarghouei, A., Breckon, T.: A comparative review of plausible hole filling strategies in the context of scene depth image completion. Computers and Graphics **72**, 39–58 (2018)
5. Atapour-Abarghouei, A., Breckon, T.: Extended patch prioritization for depth filling within constrained exemplar-based RGB-D image completion. In: Int. Conf. Image Analysis and Recognition, pp. 306–314 (2018)
6. Atapour-Abarghouei, A., Breckon, T.: Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 2800–2810 (2018)
7. Atapour-Abarghouei, A., Payen de La Garanderie, G., Breckon, T.P.: Back to butterworth - a fourier basis for 3d surface relief hole filling within rgb-d imagery. In: Int. Conf. Pattern Recognition, pp. 2813–2818. IEEE (2016)
8. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Analysis and Machine Intelligence **39**(12), 2481–2495 (2017)
9. Baek, S.H., Choi, I., Kim, M.H.: Multiview image completion with space structure propagation. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 488–496 (2016)
10. Ballester, C., Caselles, V., Verdera, J., Bertalmio, M., Sapiro, G.: A variational model for filling-in gray level and color images. In: Int. Conf. Computer Vision, vol. 1, pp. 10–16. IEEE (2001)
11. Barbero, A., Sra, S.: Fast newton-type methods for total variation regularization. In: Int. Conf. Machine Learning, pp. 313–320 (2011)
12. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.: Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graphics **28**(3), 24 (2009)
13. Berdnikov, Y., Vatolin, D.: Real-time depth map occlusion filling and scene background restoration for projected-pattern based depth cameras. In: Graphic Conf. IETP (2011)
14. Berger, K., Ruhl, K., Schroeder, Y., Bruemmer, C., Scholz, A., Magnor, M.A.: Markerless motion capture using multiple color-depth sensors. In: Vision Modeling and Visualization, pp. 317–324 (2011)
15. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Int. Conf. Computer Graphics and Interactive Techniques, pp. 417–424 (2000)
16. Bertalmio, M., Vese, L., Sapiro, G., Osher, S.: Simultaneous structure and texture image inpainting. IEEE Trans. Image Processing **12**(8), 882–889 (2003)
17. Breckon, T., Fisher, R.: Plausible 3D colour surface completion using non-parametric techniques. In: Mathematics of Surfaces XI, vol. 3604, pp. 102–120 (2005)
18. Breckon, T.P., Fisher, R.: Non-parametric 3D surface completion. In: Int. Conf. 3D Digital Imaging and Modeling, pp. 573–580 (2005)

19. Breckon, T.P., Fisher, R.: 3D surface relief completion via non-parametric techniques. IEEE Trans. Pattern Analysis and Machine Intelligence **30**(12), 2249–2255 (2008)
20. Breckon, T.P., Fisher, R.: A hierarchical extension to 3D non-parametric surface relief completion. Pattern Recognition **45**, 172–185 (2012)
21. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: Int. Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 60–65. IEEE (2005)
22. Bugeau, A., Bertalmío, M., Caselles, V., Sapiro, G.: A comprehensive framework for image inpainting. IEEE Trans. Image Processing **19**(10), 2634–2645 (2010)
23. Butler, A., Izadi, S., Hilliges, O., Molyneaux, D., Hodges, S., Kim, D.: Shake'n'sense: Reducing interference for overlapping structured light depth cameras. In: Conf. Human Factors in Computing Systems, p. 1933–1936 (2012)
24. Camplani, M., Salgado, L.: Adaptive spatiotemporal filter for low-cost camera depth maps. In: Int. Conf. Emerging Signal Processing Applications, pp. 33–36. IEEE (2012)
25. Camplani, M., Salgado, L.: Efficient spatiotemporal hole filling strategy for kinect depth maps. In: IS&T/SPIE Electronic Imaging, pp. 82,900E–82,900E (2012)
26. Cao, Y., Wu, Z., Shen, C.: Estimating depth from monocular images as classification using deep fully convolutional residual networks. IEEE Trans. Circuits and Systems for Video Technology **28**(11), 3174–3182 (2017)
27. Cavestany, P., Rodriguez, A., Martinez-Barbera, H., Breckon, T.: Improved 3d sparse maps for high-performance structure from motion with low-cost omnidirectional robots. In: Int. Conf. Image Processing, pp. 4927–4931 (2015)
28. Chan, T., Shen, J.: Mathematical models for local deterministic inpaintings. Tech. rep., Technical Report CAM TR 00-11, UCLA (2000)
29. Chang, A., Dai, A., Funkhouser, T., Halber, M., Nießner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3D: Learning from RGB-D data in indoor environments. In: Int Conf. 3D Vision (2017)
30. Chen, C., Cai, J., Zheng, J., Cham, T.J., Shi, G.: A color-guided, region-adaptive and depth-selective unified framework for kinect depth recovery. In: Int. Workshop on Multimedia Signal Processing, pp. 007–012. IEEE (2013)
31. Chen, C., Cai, J., Zheng, J., Cham, T.J., Shi, G.: Kinect depth recovery using a color-guided, region-adaptive, and depth-selective framework. ACM Trans. Intelligent Systems and Technology **6**(2), 12 (2015)
32. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. In: Advances in Neural Information Processing Systems, pp. 730–738 (2016)
33. Cong, P., Xiong, Z., Zhang, Y., Zhao, S., Wu, F.: Accurate dynamic 3d sensing with fourier-assisted phase shifting. Selected Topics in Signal Processing **9**(3), 396–408 (2015)
34. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 3213–3223 (2016)
35. Crabb, R., Tracey, C., Puranik, A., Davis, J.: Real-time foreground segmentation via range and color imaging. In: IEEE Conf. Computer Vision and Pattern Recognition Workshops, pp. 1–5 (2008)
36. Criminisi, A., Pérez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE Trans. Image Processing **13**(9), 1200–1212 (2004)
37. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3d transform-domain collaborative filtering. IEEE Trans. Image Processing **16**(8), 2080–2095 (2007)
38. Darabi, S., Shechtman, E., Barnes, C., Goldman, D.B., Sen, P.: Image melding: Combining inconsistent images using patch-based synthesis. ACM Trans. Graphics **31**(4), 82–1 (2012)
39. Daribo, I., Saito, H.: A novel inpainting-based layered depth video for 3dtv. IEEE Trans. Broadcasting **57**(2), 533–541 (2011)
40. Delage, E., Lee, H., Ng, A.Y.: A dynamic bayesian network model for autonomous 3D reconstruction from a single indoor image. In: IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 2418–2428. IEEE (2006)

41. Ding, L., Sharma, G.: Fusing structure from motion and lidar for dense accurate depth map estimation. In: Int. Conf. Acoustics, Speech and Signal Processing, pp. 1283–1287. IEEE (2017)
42. Efros, A.A., Freeman, W.T.: Image quilting for texture synthesis and transfer. In: Conf. Computer Graphics and Interactive Techniques, pp. 341–346. ACM (2001)
43. Efros, A.A., Leung, T.K.: Texture synthesis by non-parametric sampling. In: Int. Conf. Computer Vision, vol. 2, pp. 1033–1038. IEEE (1999)
44. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Int. Conf. Computer Vision, pp. 2650–2658 (2015)
45. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in Neural Information Processing Systems, pp. 2366–2374 (2014)
46. El-laithy, R.A., Huang, J., Yeh, M.: Study on the use of microsoft kinect for robotics applications. In: Position Location and Navigation Symposium, pp. 1280–1288. IEEE (2012)
47. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. Computer Vision **59**(2), 167–181 (2004)
48. Fu, D., Zhao, Y., Yu, L.: Temporal consistency enhancement on depth sequences. In: Picture Coding Symposium, pp. 342–345. IEEE (2010)
49. Gangwal, O.P., Djapic, B.: Real-time implementation of depth map post-processing for 3d-tv in dedicated hardware. In: Int. Conf. Consumer Electronics, pp. 173–174. IEEE (2010)
50. Garg, R., Carneiro, G., Reid, I.: Unsupervised CNN for single view depth estimation: Geometry to the rescue. In: Euro. Conf. Computer Vision, pp. 740–756. Springer (2016)
51. Garro, V., Mutto, C.D., Zanuttigh, P., Cortelazzo, G.M.: A novel interpolation scheme for range data with side information. In: Conf. Visual Media Production, pp. 52–60. IEEE (2009)
52. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 2414–2423 (2016)
53. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The KITTI dataset. Robotics Research pp. 1231–1237 (2013)
54. Ghiasi, G., Lee, H., Kudlur, M., Dumoulin, V., Shlens, J.: Exploring the structure of a real-time, arbitrary neural artistic stylization network. In: British Machine Vision Conference, pp. 1–12 (2017)
55. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 6602 – 6611 (2017)
56. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
57. Gortler, S.J., Grzeszczuk, R., Szeliski, R., Cohen, M.F.: The lumigraph. In: Conf. Computer Graphics and Interactive Techniques, pp. 43–54. ACM (1996)
58. Guillemot, C., Le Meur, O.: Image inpainting: Overview and recent advances. Signal Processing Magazine **31**(1), 127–144 (2014)
59. Hansard, M., Lee, S., Choi, O., Horaud, R.P.: Time-of-Flight Cameras: Principles, Methods and Applications. Springer Science & Business Media (2012)
60. Hays, J., Efros, A.A.: Scene completion using millions of photographs. ACM Trans. Graphics **26**(3), 4 (2007)
61. He, K., Sun, J., Tang, X.: Guided image filtering. In: Euro. Conf. Computer Vision, pp. 1–14. Springer (2010)
62. Heitz, G., Gould, S., Saxena, A., Koller, D.: Cascaded classification models: Combining models for holistic scene understanding. In: Advances in Neural Information Processing Systems, pp. 641–648 (2009)
63. Herrera, D., Kannala, J., Heikkilä, J., et al.: Depth map inpainting under a second-order smoothness prior. In: Scandinavian Conf. Image Analysis, pp. 555–566. Springer (2013)
64. Hervieu, A., Papadakis, N., Bugeau, A., Gargallo, P., Caselles, V.: Stereoscopic image inpainting: Distinct depth maps and images inpainting. In: Int. Conf. Pattern Recognition, pp. 4101–4104. IEEE (2010)

65. Hirschmuller, H.: Stereo processing by semi-global matching and mutual information. IEEE Trans. Pattern Analysis and Machine Intelligence **30**, 328–341 (2008)
66. Hirschmuller, H., Scharstein, D.: Evaluation of cost functions for stereo matching. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 1–8 (2007)
67. Hoiem, D., Efros, A.A., Hebert, M.: Automatic photo pop-up. In: ACM trans. Graphics, vol. 24, pp. 577–584 (2005)
68. Hoiem, D., Efros, A.A., Hebert, M.: Geometric context from a single image. In: It. Conf. Computer Vision, vol. 1, pp. 654–661. IEEE (2005)
69. Hu, J., Ozay, M., Zhang, Y., Okatani, T.: Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. arXiv preprint arXiv:1803.08673 (2018)
70. Huang, X., Wang, L., Huang, J., Li, D., Zhang, M.: A depth extraction method based on motion and geometry for 2D to 3D conversion. In: Intelligent Information Technology Application, vol. 3, pp. 294–298. IEEE (2009)
71. Ihler, A., McAllester, D.: Particle belief propagation. In: Artificial Intelligence and Statistics, pp. 256–263 (2009)
72. Ihrke, I., Kutulakos, K.N., Lensch, H., Magnor, M., Heidrich, W.: Transparent and specular object reconstruction. In: Computer Graphics Forum, vol. 29, pp. 2400–2426. Wiley Online Library (2010)
73. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Trans. Graphics **36**(4), 107 (2017)
74. Islam, A.T., Scheel, C., Pajarola, R., Staadt, O.: Robust enhancement of depth images from depth sensors. Computers & Graphics **68**, 53–65 (2017)
75. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., et al.: Kinectfusion: Real-time 3D reconstruction and interaction using a moving depth camera. In: ACM Symp. User Interface Software and Technology, pp. 559–568 (2011)
76. Jackson, P.T., Atapour-Abarghouei, A., Bonner, S., Breckon, T., Obara, B.: Style augmentation: Data augmentation via style randomization. arXiv preprint arXiv:1809.05375 pp. 1–13 (2018)
77. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in Neural Information Processing Systems, pp. 2017–2025 (2015)
78. Janarthanan, V., Jananii, G.: A detailed survey on various image inpainting techniques. Advances in Image Processing **2**(2), 1 (2012)
79. Jia, J., Tang, C.K.: Image repairing: Robust image synthesis by adaptive n-d tensor voting. In: IEEE Conf. Computer Vision and Pattern Recognition, vol. 1, pp. I–643 (2003)
80. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Euro. Conf. Computer Vision, pp. 694–711 (2016)
81. Karsch, K., Liu, C., Kang, S.B.: Depth transfer: Depth extraction from video using non-parametric sampling. IEEE Trans. Pattern Analysis and Machine Intelligence **36**(11), 2144–2158 (2014)
82. Kim, Y., Ham, B., Oh, C., Sohn, K.: Structure selective depth superresolution for rgb-d cameras. IEEE Trans. Image Processing **25**(11), 5227–5238 (2016)
83. Komodakis, N., Tziritas, G.: Image completion using efficient belief propagation via priority scheduling and dynamic pruning. IEEE Trans. Image Processing **16**(11), 2649–2661 (2007)
84. Kopf, J., Cohen, M.F., Lischinski, D., Uyttendaele, M.: Joint bilateral upsampling. ACM Trans. Graphics **26**(3), 96 (2007)
85. Kumar, V., Mukherjee, J., Mandal, S.K.D.: Image inpainting through metric labeling via guided patch mixing. IEEE Trans. Image Processing **25**(11), 5212–5226 (2016)
86. Kuznietsov, Y., Stückler, J., Leibe, B.: Semi-supervised deep learning for monocular depth map prediction. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 6647–6655 (2017)
87. Ladicky, L., Shi, J., Pollefeys, M.: Pulling things out of perspective. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 89–96 (2014)

88. Lai, K., Bo, L., Ren, X., Fox, D.: A large-scale hierarchical multi-view rgb-d object dataset. In: Int. Conf. Robotics and Automation, pp. 1817–1824. IEEE (2011)
89. Lai, P., Tian, D., Lopez, P.: Depth map processing with iterative joint multilateral filtering. In: Picture Coding Symposium, pp. 9–12. IEEE (2010)
90. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: Int. Conf. 3D Vision, pp. 239–248. IEEE (2016)
91. Lee, J.H., Choi, I., Kim, M.H.: Laplacian patch-based image synthesis. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 2727–2735 (2016)
92. Lee, S.B., Ho, Y.S.: Discontinuity-adaptive depth map filtering for 3d view generation. In: Int. Conf. Immersive Telecommunications, p. 8. ICST (2009)
93. Levin, A., Lischinski, D., Weiss, Y.: A closed-form solution to natural image matting. IEEE Trans. Pattern Analysis and Machine Intelligence **30**(2), 228–242 (2008)
94. Li, B., Dai, Y., He, M.: Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference. Pattern Recognition (2018)
95. Li, C., Wand, M.: Combining markov random fields and convolutional neural networks for image synthesis. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 2479–2486 (2016)
96. Lindner, M., Schiller, I., Kolb, A., Koch, R.: Time-of-flight sensor calibration for accurate range sensing. Computer Vision and Image Understanding **114**(12), 1318–1328 (2010)
97. Liu, B., Gould, S., Koller, D.: Single image depth estimation from predicted semantic labels. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 1253–1260. IEEE (2010)
98. Liu, F., Shen, C., Lin, G.: Deep convolutional neural fields for depth estimation from a single image. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 5162–5170 (2015)
99. Liu, F., Shen, C., Lin, G., Reid, I.: Learning depth from single monocular images using deep convolutional neural fields. IEEE Trans. Pattern Analysis and Machine Intelligence **38**(10), 2024–2039 (2016)
100. Liu, J., Gong, X., Liu, J.: Guided inpainting and filtering for kinect depth maps. In: Int. Conf. Pattern Recognition, pp. 2055–2058. IEEE (2012)
101. Liu, M., Salzmann, M., He, X.: Discrete-continuous depth estimation from a single image. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 716–723 (2014)
102. Liu, S., Lai, P., Tian, D., Gomila, C., Chen, C.W.: Joint trilateral filtering for depth map compression. In: Visual Communications and Image Processing, pp. 77,440F–77,440F. International Society for Optics and Photonics (2010)
103. Liu, S., Wang, Y., Wang, J., Wang, H., Zhang, J., Pan, C.: Kinect depth restoration via energy minimization with tv 21 regularization. In: Int. Conf. Image Processing, pp. 724–724. IEEE (2013)
104. Lu, S., Ren, X., Liu, F.: Depth enhancement via low-rank matrix completion. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 3390–3397 (2014)
105. Ma, Y., Worrall, S., Kondoz, A.M.: Automatic video object segmentation using depth information and an active contour model. In: Workshop on Multimedia Signal Processing, pp. 910–914. IEEE (2008)
106. Matyunin, S., Vatolin, D., Berdnikov, Y., Smirnov, M.: Temporal filtering for depth maps generated by kinect depth camera. In: 3DTV Conference, pp. 1–4. IEEE (2011)
107. Miao, D., Fu, J., Lu, Y., Li, S., Chen, C.W.: Texture-assisted kinect depth inpainting. In: Int. Symp. Circuits and Systems, pp. 604–607. IEEE (2012)
108. Min, D., Lu, J., Do, M.N.: Depth video enhancement based on weighted mode filtering. IEEE Trans. Image Processing **21**(3), 1176–1190 (2012)
109. Mueller, M., Zilly, F., Kauff, P.: Adaptive cross-trilateral depth map filtering. In: 3DTV Conference, pp. 1–4. IEEE (2010)
110. Nguyen, H.T., Do, M.N.: Image-based rendering with depth information using the propagation algorithm. In: Int. Conf. Acoustics, Speech, and Signal Processing, pp. 589–592 (2005)
111. Nguyen, K., Fookes, C., Sridharan, S., Tistarelli, M., Nixon, M.: Super-resolution for biometrics: A comprehensive survey. Pattern Recognition **78**, 23–42 (2018)

112. Nguyen, Q.H., Do, M.N., Patel, S.J.: Depth image-based rendering from multiple cameras with 3d propagation algorithm. In: Int. Conf. Immersive Telecommunications, p. 6. ICST (2009)
113. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 2536–2544 (2016)
114. Peng, J., Hazan, T., McAllester, D., Urtasun, R.: Convex max-product algorithms for continuous MRFs with applications to protein folding. In: Int. Conf. Machine Learning, pp. 729–736 (2011)
115. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. IEEE Trans. Pattern Analysis and Machine Intelligence **12**(7), 629–639 (1990)
116. Petschnigg, G., Szeliski, R., Agrawala, M., Cohen, M., Hoppe, H., Toyama, K.: Digital photography with flash and no-flash image pairs. In: ACM Trans. Graphics, vol. 23, pp. 664–672 (2004)
117. Po, L.M., Zhang, S., Xu, X., Zhu, Y.: A new multi-directional extrapolation hole-filling method for depth-image-based rendering. In: Int. Conf. Image Processing, pp. 2589–2592. IEEE (2011)
118. Popat, K., Picard, R.W.: Novel cluster-based probability model for texture synthesis, classification, and compression. In: Visual Communications, pp. 756–768 (1993)
119. Pritch, Y., Kav-Venaki, E., Peleg, S.: Shift-map image editing. In: Int. Conf. Computer Vision, vol. 9, pp. 151–158 (2009)
120. Qi, F., Han, J., Wang, P., Shi, G., Li, F.: Structure guided fusion for depth map inpainting. Pattern Recognition Letters **34**(1), 70–76 (2013)
121. Richard, M.M.O.B.B., Chang, M.Y.S.: Fast digital image inpainting. In: Int. Conf. Visualization, Imaging and Image Processing, pp. 106–107 (2001)
122. Richardt, C., Stoll, C., Dodgson, N.A., Seidel, H.P., Theobalt, C.: Coherent spatiotemporal filtering, upsampling and rendering of rgbz videos. In: Computer Graphics Forum, vol. 31, pp. 247–256. Wiley Online Library (2012)
123. Ringbeck, T., Möller, T., Hagebeuker, B.: Multidimensional measurement by using 3D PMD sensors. Advances in Radio Science **5**, 135 (2007)
124. Rousseeuw, P.J.: Least median of squares regression. American Statistical Association **79**(388), 871–880 (1984)
125. Sabov, A., Krüger, J.: Identification and correction of flying pixels in range camera data. In: Conf. Computer Graphics, pp. 135–142. ACM (2008)
126. Sarbolandi, H., Lefloch, D., Kolb, A.: Kinect range sensing: Structured-light versus time-of-flight kinect. Computer Vision and Image Understanding **139**, 1–20 (2015)
127. Saxena, A., Chung, S.H., Ng, A.Y.: Learning depth from single monocular images. In: Advances in Neural Information Processing Systems, pp. 1161–1168 (2006)
128. Saxena, A., Sun, M., Ng, A.Y.: Make3d: Learning 3D scene structure from a single still image. IEEE Trans. Pattern Analysis and Machine Intelligence **31**(5), 824–840 (2009)
129. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. Computer Vision **47**, 7–42 (2002)
130. Sheng, L., Ngan, K.N., Li, S.: Temporal depth video enhancement based on intrinsic static structure. In: Int. Conf. Image Processing, pp. 2893–2897. IEEE (2014)
131. Suthar, R., Patel, M.K.R.: A survey on various image inpainting techniques to restore image. Int. J. Engineering Research and Applications **4**(2), 85–88 (2014)
132. Tao, M.W., Srinivasan, P.P., Malik, J., Rusinkiewicz, S., Ramamoorthi, R.: Depth from shading, defocus, and correspondence using light-field angular coherence. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 1940–1948 (2015)
133. Telea, A.: An image inpainting technique based on the fast marching method. Graphics Tools **9**(1), 23–34 (2004)
134. Tippetts, B., Lee, D.J., Lillywhite, K., Archibald, J.: Review of stereo vision algorithms and their suitability for resource-limited systems. Real-Time Image Processing **11**(1), 5–25 (2016)

135. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: Int. Conf. Computer Vision, pp. 839–846. IEEE (1998)
136. Vijayanagar, K.R., Loghman, M., Kim, J.: Real-time refinement of kinect depth maps using multi-resolution anisotropic diffusion. Mobile Networks and Applications **19**(3), 414–425 (2014)
137. Wang, J., An, P., Zuo, Y., You, Z., Zhang, Z.: High accuracy hole filling for kinect depth maps. In: SPIE/COS Photonics Asia, pp. 92,732L–92,732L (2014)
138. Wang, L., Huang, Z., Gong, Y., Pan, C.: Ensemble based deep networks for image super-resolution. Pattern Recognition **68**, 191–198 (2017)
139. Wei, L.Y., Lefebvre, S., Kwatra, V., Turk, G.: State of the art in example-based texture synthesis. In: Eurographics State of the Art Report, pp. 93–117 (2009)
140. Wexler, Y., Shechtman, E., Irani, M.: Space-time completion of video. IEEE Trans. Pattern Analysis and Machine Intelligence **29**(3), 463–476 (2007)
141. Whyte, O., Sivic, J., Zisserman, A.: Get out of my picture! internet-based inpainting. In: British Machine Vision Conference, pp. 1–11 (2009)
142. Wu, Y., Ying, S., Zheng, L.: Size-to-depth: A new perspective for single image depth estimation. arXiv preprint arXiv:1801.04461 (2018)
143. Xie, J., Girshick, R., Farhadi, A.: Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks. In: Euro. Conf. Computer Vision, pp. 842–857. Springer (2016)
144. Xu, X., Po, L.M., Cheung, C.H., Feng, L., Ng, K.H., Cheung, K.W.: Depth-aided exemplar-based hole filling for dibr view synthesis. In: Int. Symp. Circuits and Systems, pp. 2840–2843. IEEE (2013)
145. Xue, H., Zhang, S., Cai, D.: Depth image inpainting: Improving low rank matrix completion with low gradient regularization. IEEE Trans. Image Processing **26**(9), 4311–4320 (2017)
146. Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H.: High-resolution image inpainting using multi-scale neural patch synthesis. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 4076–4084 (2017)
147. Yang, J., Ye, X., Li, K., Hou, C., Wang, Y.: Color-guided depth recovery from rgb-d data using an adaptive autoregressive model. IEEE Trans. Image Processing **23**(8), 3443–3458 (2014)
148. Yang, N.E., Kim, Y.G., Park, R.H.: Depth hole filling using the depth distribution of neighboring regions of depth holes in the kinect sensor. In: Int. Conf. Signal Processing, Communication and Computing, pp. 658–661. IEEE (2012)
149. Yang, Q., Tan, K.H., Culbertson, B., Apostolopoulos, J.: Fusion of active and passive sensors for fast 3d capture. In: Int. Workshop on Multimedia Signal Processing, pp. 69–74. IEEE (2010)
150. Yeh*, R.A., Chen*, C., Lim, T.Y., G., S.A., Hasegawa-Johnson, M., Do, M.N.: Semantic image inpainting with deep generative models. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 6882–6890 (2017)
151. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 1–15 (2018)
152. Zhan, H., Garg, R., Weerasekera, C.S., Li, K., Agarwal, H., Reid, I.: Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 340–349 (2018)
153. Zhang, G., Jia, J., Hua, W., Bao, H.: Robust bilayer segmentation and motion/depth estimation with a handheld camera. IEEE Trans. Pattern Analysis and Machine Intelligence **33**(3), 603–617 (2011)
154. Zhang, L., Shen, P., Zhang, S., Song, J., Zhu, G.: Depth enhancement with improved exemplar-based inpainting and joint trilateral guided filtering. In: Int. Conf. Image Processing, pp. 4102–4106. IEEE (2016)
155. Zhang, L., Tam, W.J., Wang, D.: Stereoscopic image generation based on depth images. In: Int. Conf. Image Processing, vol. 5, pp. 2993–2996. IEEE (2004)

156. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: Euro. Conf. Computer Vision, pp. 649–666 (2016)
157. Zhang, Y., Funkhouser, T.: Deep depth completion of a single RGB-D image. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 175–185 (2018)
158. Zhang, Z.: Microsoft kinect sensor and its effect. IEEE Multimedia **19**(2), 4–10 (2012)
159. Zheng, C., Cham, T.J., Cai, J.: T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks. In: Euro. Conf. Computer Vision, pp. 767–783 (2018)
160. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 6612–6619 (2017)
161. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. Int. Conf. Computer Vision pp. 2242 – 2251 (2017)
162. Zhuo, W., Salzmann, M., He, X., Liu, M.: Indoor scene structure analysis for single image depth estimation. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 614–622 (2015)

Index